WHAT IS CLAIMED IS:

1. A method for determining a value representing a difference between a first record comprising a first plurality of data fields and a second record comprising a second plurality of data fields, each of the first plurality of data fields corresponding to a respective one of the second plurality of data fields, the method comprising:

for each of the first plurality of data fields, determining a first value representing a difference between data specified in the data field and data specified in a respective one of the second plurality of data fields;

for each of the second plurality of data fields, determining a second value representing a difference between data specified in the data field and data specified in a respective one of the first plurality of data fields; and

determining a third value representing a difference between the first record and the second record based on the determined first and second values.

2. A method according to Claim 1, wherein the step of determining the third value comprises:

determining, for each of the first plurality of data fields and respective ones of the second plurality of data fields, a fourth value based on a mean of a first value determined for one of the first plurality of data fields and a second value determined for a respective one of the second plurality of data fields; and

summing the determined fourth values.

3. A method according to Claim 1, wherein the step of determining the third value comprises:

determining a sum of the determined first values and the determined second values; and

dividing the sum by two.

4. A method according to Claim 1, wherein the step of determining the first value and the step of determining the second value comprise identical steps performed with respect to different inputs.

5. A method according to Claim 1, wherein the step of determining the first value comprises:

determining an asymmetric spelling distance as a normalized cost for converting first input data to second input data via a sequence of operations; and

wherein the step of determining the second value comprises:

determining an asymmetric spelling distance as a normalized cost for converting second input data to first input data via the sequence of operations

6. A method according to Claim 5, wherein, in the step of determining the first value, the first input data is data specified in one of the first plurality of data fields and the second input data is data specified in a respective one of the second plurality of data fields, and

wherein, in the step of determining the second value, the first input data is data specified in one of the second plurality of data fields and the second input data is data specified in a respective one of the first plurality of data fields.

7. A method according to Claim 1, further comprising:

converting numerical data specified in the one or more of the first plurality of data fields and the second plurality of data fields to text data.

8. A method according to Claim 1, wherein the first plurality of data fields and the second plurality of data fields include only those fields of the first record and the second record that specify data that is not identical to data specified in a respective field.

9. A method for use in loading data in a data warehouse, comprising:

receiving a plurality of records, each of the plurality of records including a plurality of data fields;

identifying a plurality of groups of records, wherein data specified in one or more of the plurality of data fields included in a record of a group is identical to data

specified in one or more corresponding data fields included in each other record of the group;

determining, for each group, values representing differences between each record of a group and each other record of the group; and

identifying at least two of the plurality records as duplicates based on a determined value representing a difference between the two records.

10. A method according to Claim 9, wherein the step of determining values comprises:

for each of a first plurality of data fields of a first record, determining a first value representing a difference between data specified in the data field and data specified in a respective one of a second plurality of data fields of a second record;

for each of the second plurality of data fields, determining a second value representing a difference between data specified in the data field and data specified in a respective one of the first plurality of data fields; and

determining a third value representing a difference between the first record and the second record based on the determined first and second values.

11. A method according to Claim 10, wherein the first plurality of data fields and the second plurality of data fields do not include the one or more corresponding data fields specifying identical data in each record.

12. A method according to Claim 9, further comprising:

receiving identification of the one or more of the plurality of data fields from a user.

13. A method according to Claim 9, further comprising:

formatting the received records based on a standard format for data specified in each of the plurality of data fields.

14. A method according to Claim 9, further comprising:

identifying one or more hoax records,

wherein the identified one or more hoax records are not included in any of the plurality of groups of records.

15. A method according to Claim 9, further comprising:

identifying a first record and a second record of a group of records in which data specified in all of the plurality of data fields of the first record is identical to data specified in all of the plurality of data fields of the second record,

wherein the identified second record is not included in any of the plurality of groups of records.

16. A method according to Claim 15, further comprising:

storing the second record in the data warehouse in association with an identifier identical to an identifier associated with the first record.

17. A method according to Claim 9, further comprising:

identifying a first record and a second record of a group of records as duplicates based on business rules,

wherein the second record is not included in any of the plurality of groups of records.

18. A method according to Claim 17, further comprising:

storing the second record in the data warehouse in association with an identifier identical to an identifier associated with the first record.

19. A method according to Claim 9, the identifying step comprising:

determining that the value representing the difference between the two records is below a threshold value.

20. A method according to Claim 9, the identifying step comprising:

determining that the value representing the difference between the two records is within a specified range of values;

presenting the two records to a user; and

receiving an indication from the user that the two records are duplicate records.

21. A method according to Claim 20, further comprising:

storing one of the two records in the data warehouse in association with an identifier identical to an identifier associated with the other of the two records.

22. A method for loading data in a data warehouse storing existing records, comprising:

receiving a plurality of new records;

for each of the plurality of new records, determining values representing differences between a new record and one or more of the existing records;

identifying at least one of the plurality of new records and one of the existing records as duplicates based on a determined value representing a difference between the two records; and

storing the at least one of the plurality of new records in the data warehouse in association with an identifier identical to an identifier associated with the one of the existing records.

23. A method according to Claim 22, wherein, in the determining step, the one or more of the existing records comprise all of the existing records.

24. A method according to Claim 22, wherein, in the determining step, the one or more of the existing records comprise only the existing records of which data specified in particular fields is identical to data specified in corresponding fields of the new record.

25. A method according to Claim 22, wherein the step of determining values comprises:

for each of a first plurality of data fields of the new record, determining a first value representing a difference between data specified in the data field and data

specified in a respective one of a second plurality of data fields of one of the one or more existing records;

for each of the second plurality of data fields, determining a second value representing a difference between data specified in the data field and data specified in a respective one of the first plurality of data fields; and

determining a third value representing a difference between the new record and the one of the one or more existing records based on the determined first and second values.

26. A method according to Claim 25, wherein the first plurality of data fields and the second plurality of data fields include only those fields of the new record and the one of the one or more existing records that specify data that is not identical to data specified in a respective field.

27. A method for loading data in a data warehouse, comprising:

receiving a plurality of records;

for each of the plurality of records, determining values representing differences between a record and each other of the plurality of records;

identifying at least two of the plurality records as duplicates based on a determined value representing a difference between the two records; and

storing the two records in the data warehouse in association with a same identifier.

28. A method according to Claim 27, wherein the step of determining values comprises:

for each of a first plurality of data fields of the record, determining a first value representing a difference between data specified in the data field and data specified in a respective one of a second plurality of data fields of one of the other of the plurality of records;

for each of the second plurality of data fields, determining a second value representing a difference between data specified in the data field and data specified in a respective one of the first plurality of data fields; and

determining a third value representing a difference between the record and the one of the other of the plurality of records based on the determined first and second values.

29. A method according to Claim 28, wherein the first plurality of data fields and the second plurality of data fields include only those fields of the record and the one of the other of the plurality of records that specify data that is not identical to data specified in a respective field.

30. A system for storing data, comprising:

a device for transmitting a plurality of new records; and

a data warehouse for storing existing records, for receiving the transmitted plurality of records, for determining values representing differences between a new record and one or more of the existing records for each of the plurality of new records, for identifying at least one of the plurality of new records and one of the existing records as duplicates based on a determined value representing a difference between the two records, and for storing the at least one of the plurality of new records in association with an identifier identical to an identifier associated with the one of the existing records.

31. A system according to Claim 30, wherein the data warehouse determines, for each of a first plurality of data fields of the record, a first value representing a difference between data specified in the data field and data specified in a respective one of a second plurality of data fields of one of the other of the plurality of records,

determines, for each of the second plurality of data fields, a second value representing a difference between data specified in the data field and data specified in a respective one of the first plurality of data fields, and

determines a third value representing a difference between the record and the one of the other of the plurality of records based on the determined first and second values.

32. A computer-readable medium storing processor-executable process steps to determine a value representing a difference between a first record comprising a first plurality of data fields and a second record comprising a second plurality of data fields, each of the first plurality of data fields corresponding to a respective one of the second plurality of data fields, the steps comprising:

a step to determine, for each of the first plurality of data fields, a first value representing a difference between data specified in the data field and data specified in a respective one of the second plurality of data fields;

a step to determine, for each of the second plurality of data fields, a second value representing a difference between data specified in the data field and data specified in a respective one of the first plurality of data fields; and

a step to determine a third value representing a difference between the first record and the second record based on the determined first and second values.

33. A medium according to Claim 32, wherein the step to determine the first value and the step to determine the second value comprise identical steps performed with respect to different inputs.

34. A medium according to Claim 32, wherein the step to determine the first value comprises:

a step to determine an asymmetric spelling distance as a normalized cost for converting first input data to second input data via a sequence of operations; and

wherein the step to determine the second value comprises:

a step to determine an asymmetric spelling distance as a normalized cost for converting second input data to first input data via the sequence of operations.

35. A medium according to Claim 34, wherein, in the step to determine the first value, the first input data is data specified in one of the first plurality of data fields and the second input data is data specified in a respective one of the second plurality of data fields, and

wherein, in the step to determine the second value, the first input data is data specified in one of the second plurality of data fields and the second input data is data specified in a respective one of the first plurality of data fields.

36. A medium according to Claim 32, wherein the first plurality of data fields and the second plurality of data fields include only those fields of the first record and the second record that specify data that is not identical to data specified in a respective field.

37. A data warehouse, comprising:

a processor; and

a storage device in communication with the processor and storing instructions adapted to be executed by the processor to:

receive a plurality of records;

determine values representing differences between a new record and one or more of the existing records for each of the plurality of new records;

identify at least one of the plurality of new records and one of the existing records as duplicates based on a determined value representing a difference between the two records; and

store the at least one of the plurality of new records in association with an identifier identical to an identifier associated with the one of the existing records.

38. A data warehouse, comprising:

a processor; and

a storage device in communication with the processor and storing instructions adapted to be executed by the processor to:

determine, for each of a first plurality of data fields of a first record, a first value representing a difference between data specified in the data field and data specified in a respective one of a second plurality of data fields of a second record,

determine, for each of the second plurality of data fields, a second value representing a difference between data specified in the data field and data specified in a respective one of the first plurality of data fields, and

determine a third value representing a difference between the first record and the second record based on the determined first and second values.

39. A data warehouse according to Claim 38, wherein the instructions adapted to be executed by the processor to determine the first value and to determine the second value comprise identical steps performed with respect to different inputs.

40. A data warehouse according to Claim 38, wherein the instructions adapted to be executed by the processor to determine the first value comprise instructions adapted to be executed by the processor to:

determine an asymmetric spelling distance as a normalized cost for converting first input data to second input data via a sequence of operations; and

wherein the instructions adapted to be executed by the processor to determine the second value comprise instructions adapted to be executed by the processor to:

determine an asymmetric spelling distance as a normalized cost for converting the second input data to the first input data via the sequence of operations.